CHEMINGUI Mohamed*, BAHSOUN Wahiba**
*** mohamedchemingui@gmail.com: ISAMM, RIADI**
****wahiba.bahsoun@irit.fr:IRIT**

Big data Study and Manipulating: Experience in industrial context

Abstract: **This article presents a summary of the internship graduation project at the Higher Institute of Multimedia Arts of Manouba. This project took place at the Institute for Computer Science of Toulouse (IRIT), conducted jointly with the Laboratory RIADI. Its objective is "the study, Bulky Handling and Data Processing in an Industrial context.**

The context of this project is to migrate the massive data stored in a traditional format (.csv, .txt, ...) to NoSQL DBMS in Big Data solutions to facilitate their treatment. This work was performed on the Oracle Big Data solution.

Key words: Big Data, NoSQL, JSON.

I. INTRODUCTION

In recent years, the socio-economic background becomes aware of the exponential growth of its data and organizes to develop tools, methods, techniques and procedures to find solutions to this problem. How, even these companies will use the powers of the "Data Scientist" to be able to address their specific needs such as collection, storage and analysis of data.
Actually, software companies offer a family of tools that tend to answer some of Big data characteristics that we define far more such as volume, variability, velocity and veracity of data.
Several new tools were presented by various software vendors. The first important tool is the Database Management System (DBMS) Not only SQL NoSQL offering new techniques for data management. The second tool, Hadoop, which is a very powerful Framework which addresses several aspects of Big Data and finally Apache Spark which is a data analysis Framework in the open source operating environment designed for speed and ease of use tools.

II. PR OBLEMATIC

Today one of the major problems of businesses, is to anticipate the needs of their customers and offer them better service, because the data volume continues to evolve and increase. This is not possible with conventional tools. Their challenge is to improve storage efficiency and facilitate access to data for analytical needs.

The problem of this subject is to understand the service data, control their life and to propose ways to exploit them in the Oracle environment already installed at the customer, without changing user habits.

We had to propose an approach for data transfer, organize, store and finally facilitate access to the data for statistical processing.

In this context, the company needs to bring together all of its data in a global database. The volume of handled data is about one Tera bytes. This led the company to make the choice of Oracle Big Data solutions that offer the necessary storage space and computing power to handle this large volume.

Our work was performed on the Oracle Big Data Appliance to use the Oracle NoSQL DBMS for data storage, to also use the Spark and Hadoop technologies for processing data.

III. Big Data

In what follows, we define the concept of Big Data in accordance. Next, we present the different technologies mentioned above, related to Big Data such as NoSQL, Hadoop and Spark, with a particular emphasis on the Oracle Big Data architecture.

A. Definition

Before you define big data, it is essential to say what it is not “Big Data is NOT a bigger data warehouse” [1]. In other words, Big Data, not data centers increasingly large to always store more data. “Big Data is a set of more or less structured data that are becoming so large and difficult to treat with conventional tool database management” [2]. “The Big data encompasses a set of technologies and practices for storing very large amounts of data and analyze very quickly” [3].

B. Big Data characteristics

In 2012, the American company **Gartner** has laid the foundation for the definition of Big Data, about **4V** [4]:

Volume*: Treat a data amount increasingly important.*

Velocity *(speed): Process and analyze all of this data in a limited time.*

Variety of data*: Treat renormalized data, unstructured or semi-structured data.*

Veracity*: confidence in data*

IV. SGBD « NoSQL »

A. Definition

NoSQL means "Not Only SQL". This term refers to all databases that are opposed to the concept of relational DBMS. Indeed, NoSQL comes not replace relational databases but offer an alternative or supplement the functionality of the RDBMS to provide more interesting solutions in some contexts [5] [6].

B. NoSQL characteristics

NoSQL databases also respond to the theorem of Eric Brewer's CAP that is better adapted to distributed systems. This theorem states that any distributed system can meet the following requirements [6] [7]:NoSQL databases also respond to the theorem of Eric Brewer's CAP that is better adapted to distributed systems. This theorem states that any distributed system can meet the following requirements [6] [7]:

Consistency: all nodes in the system exactly see the same data at the same time.

High availability: the loss of nodes does not prevent the survivors to continue to function properly, the data remains accessible.

Partitioning tolerance: the system is partitioned, no less than a total break down of the network should prevent it from responding properly.

V. Oracle NoSQL Data base

A. Description

The Oracle NoSQL database based on the model key / value. It provides a distributed storage pairs that offers throughput and scalable performance. In other words, it responds to network requests to store and retrieve data that is organized in key-value pairs. Among its unique features, the Oracle NoSQL DBMS provides major and minor keys [5] [8].

This NoSQL DBMS provides all operations create, read, modify and delete "CRUD" (Create, Read, Update and Delete) with guarantees of sustainability Instantiable. It is intended for any application that requires data to key / value pairs, each pair is placed on several machines based on the result of a hashing function on the key [9].

In particular, the key-value pair will be placed on a single master node and a configurable number of replica nodes. All write operations and update for a key-value pair go to the master node for this first pair. This replication is generally asynchronously.

B. AVSC File « AVRO SHEMA »

Avro is used to define the data schema to be stored in the Oracle NoSQL database. This schema describes the fields in the allowed value, with their data types.

The Avro API is the result of an open source project provided by the Apache Software Foundation. Using Avro schemas allows records to be stored in the Oracle NoSQL database by a key / value pair in a binary format. To apply the schema to the part value of a record you must use an Avro Binding. Avro Binding is used to serialize values in a binary format before writing, and de-serialization values after reading

C. JSON

JSON (JavaScript Object Notation – Notation Objet issue de JavaScript) is a lightweight data exchange format.. It is easy to read and write for humans. It is easily analyzable or generable by machines. It is based on a subset of the JavaScript programming language. It is based was subset of the JavaScript programming language, but uses conventions that

are familiar to any programmer familiar with C descendant languages, for exemple : C itself, C++, C#, Java, JavaScript, Perl, Python and many others. These properties make JSON an ideal data exchange language. [10]
We can point out that at present, this language is very coveted by training for use in projects and Industrial Research.

VI. R LANGAGE (programming language and environment Statistics)

R is an environment free of statistical and graphical calculations. It provides, also, a programming language, a high level of graphics and interfaces with other languages. [11]

R is supplied with the Oracle Big Data solution, it was adopted in the environment for use by statisticians profiles, including data analysts, the data the scientists and the Data Scientist in performing advanced statistical analysis on data. It generates sophisticated graphics [12].
According to surveys, in recent years has become very popular R.

VII. CONTRIBUTION

The solution is to realize a data processing environment in order to convert them into a data format that supports querying data and statistical analysis.
The first phase is study data which consists to collect and to analyze the sources of data received.
The second is to manipulate the data discussed in the previous phase.

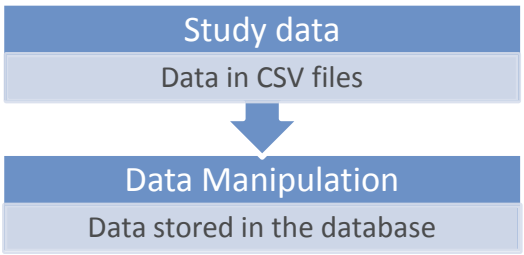

Figure 1 : Process data studies

A. Study data

The study data phase which consists to collect and to analyze the sources of data sources provided in order to understand their nature and structure.
Studied the files do not have the same structure, and are composed of more parameters and each parameter has a semantic meaning in a particular context.
Moreover, this difficulty, it was necessary to extract relevant and interesting data for a specific duration in time, this led us to conclude that each parameter has a temporal status.
The help data users has enabled us to overcome some data comprehension difficulties to establish a true diagnosis to extract interesting data for a given duration.
In particular, parameters with missing data.
The processing of such data with missing values (N/A) is a concrete problem. Because the unavailability of a value may be an information processing and analyzing, for example it could be related to an event or a specific abnormality in the company product.
To solve this problem, the missing values must not be removed and replaced while keeping their specificity.
To solve this problem, the missing values must not be removed and replaced while keeping their specificity.
To replace missing values we completed the replacement of (n / a) using an interpolation method "na.interp" using the package "forecast" with the language R.
This method allows to interpolate the missing values in a time series.
These treatments have helped us understand the data, each parameter has a semantic sense, this study has allowed us to understand all the parameters manipulated for a concept.

B. Data Processing

This phase relies on the data processing phase, which aims to transform data into JSON format for storage in the Oracle NoSQL database.
The manipulation is done according to the following two steps:

- Transformation of the CSV file in JSON
- Storage JSON file in the database NoSQL

1) Transformation

The data is in CSV format. To store data in the NoSQL database, we had to convert them into an extension NoSQL "AVRO, JSON ...»
We have chosen to convert to JSON format, which responds most to our needs.

To make the conversion, you must read the data from the CSV file and convert them to JSON file. NoSQL database storage requires the presence of a data description file with ".avsc" extension.

This file is used to describe the JSON file structure and the variables types, it is used in order to perform specific queries and facilitate data manipulation.
For this, we developed a program that transforms the data into JSON format and dynamically generates a description file with ".avsc" extension.

Practically, this new program is developed with the Python language. This program takes in input CSV file. In output, the program generates a JSON file format and its file description which may be called data schema.

Our program detects the number of columns and the typing mode to dynamically generate the data pattern as shown in Figure 2.

```
{
    "type": "record",
    "name": "Trace_vol-test",
    "namespace": "airbus",
    "fields": [
        {"name": "PARAM1", "type": "double","default":"0"},
        {"name": "PARAM2", "type": "string","default":"NONE"}
        {"name": "PARAM3", "type": "double","default":"0"},
        {"name": "PARAM4", "type": "double","default":"0"},
        {"name": "PARAM5", "type": "int","default":-1},
        {"name": "PARAM6", "type": "int","default":-1},
        {"name": "PARAM7", "type": "int","default":-1},
        {"name": "PARAM8", "type": "double","default":"0"},
        {"name": "PARAM9", "type": "double","default":"0"},
        {"name": "PARAM10", "type": "double","default":"0"},
        {"name": "PARAM11", "type": "double","default":"0"},
```

Figure 2 : Data Schema « File.avsc ».

2) Storage

The storage part includes data storing in the NoSQL database. It is performed in two steps:

- Adding schema
- Data storage

a) Adding schema

To store data, it is necessary to add the data schema in the store to describe the JSON file structure.

The use of schema allows records to be stored in the Oracle NoSQL database by a key / value pair in a binary format.

b) Data storage

Data is organized in key-value records form for storage in the NoSQL database.

The first part of the key storage is a composed key of a major key and a minor key.

So the major key contains the flight number and the seconds number since the beginning of the year and the minor key contains the identifier of each parameter.

The value is the second part of the record that must contain the value of the sensor to store.

In practice we used a Java class to store data. First we must identify the schema in the data warehouse and then save the data using the driver for the Oracle NoSQL database and a set of Java APIs.

After storing the data in this way, handling becomes easy due to the file description. Below, Figure 3 shows the result of storing records in the database NoSQL.

```
/voltest/PARAM1/1.4547463982508E7/-/PARAM10
"1.04"
/voltest/PARAM1/1.4547463982508E7/-/PARAM11
"2.491414"
/voltest/PARAM1/1.4547463982508E7/-/PARAM12
"4.596985"
/voltest/PARAM1/1.4547463982508E7/-/PARAM13
"1052.630493"
/voltest/PARAM1/1.4547463982508E7/-/PARAM14
"150.328445"
/voltest/PARAM1/1.4547463982508E7/-/PARAM15
"18.0"
/voltest/PARAM1/1.4547463982508E7/-/PARAM16
"16.0"
/voltest/PARAM1/1.4547463982508E7/-/PARAM17
"16.0"
/voltest/PARAM1/1.4547463982508E7/-/PARAM18
"16.0"
/voltest/PARAM1/1.4547463982508E7/-/PARAM19
"18.0"
/voltest/PARAM1/1.4547463982508E7/-/PARAM2
"169-08:57:43-982.508"
/voltest/PARAM1/1.4547463982508E7/-/PARAM20
"14.0"
```

Figure 3: Storing data in the database

C. Intégration

This phase consists of integrating all the modules of our project
Below, Figure 4 presents the integrated system architecture mode.

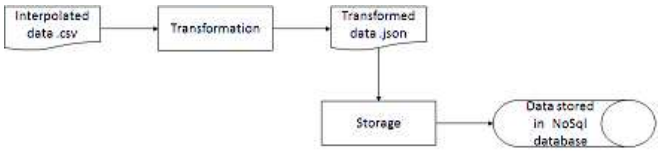

Figure 4: System Architecture

Firstly, NoSQL database storage requires data conversion into a format recognized by the base, in our case it is JSON format.

We stored the JSON data in the NoSQL database in the form of key-value pairs along with their description files.

VIII. CONCLUSION

Big Data technologies are now booming. In the coming years, these technologies will be increasingly used to solve new problems for the management, processing and analysis of data.

It is in this context that fits our work done within RIADI and IRIT laboratories. The challenge of course is to provide a system of study and manipulation of data in accordance with the characteristics of Big Data.

IX. REFERENCES:

[1] P. Doscher, "LucidWorks," April 12, 2012.

[2] F. BERNAGER, "Big Data Analyse et valorisation de masses de données".

[3] http://www.piloter.org/business-intelligence/big-data-definition.htm.

[4] http://www.gartner.com/it-glossary/big-data.

[5] L. C. P. A. R. T. F. BUGIOTTI, "Database Design for NoSQL Systems chez HAL," Roma, 2014.

[6] http://blog.neoxia.com/nosql-5-minutes-pour-comprendre/.

[7] http://blog.yocto.re/sql-or-nosql/.

[8] G. S. w. N. D. Oracle.

[9] S. H. C. L. A. Josh, "Oracle NoSQL Database – Scalable, Transactional Key-value Store,» chez The Second International Conference on Advances in Information Mining and Management," 2012.

[10]JSON GROUP: http://www.json.org/json-fr.html

[11] manuel R version 3.2.2 (2015-08-14). https://cran.r-project.org/doc/manuals/r-release/R-lang.pdf

[12] ORACLE http://www.oracle.com/technetwork/topics/bigdata/r-offerings-1566363.html